We Could, but Should We? Ethical Considerations for Providing Access to GeoCities and Other Historical Digital Collections

Jimmy Lin,[1] Ian Milligan,[2] Douglas W. Oard,[3,4] Nick Ruest,[5] Katie Shilton[3]
[1] David R. Cheriton School of Computer Science, University of Waterloo
[2] Department of History, University of Waterloo
[3] College of Information Studies and [4] UMIACS, University of Maryland, College Park
[5] York University Libraries

ABSTRACT

We live in an era in which the ways that we can make sense of our past are evolving as more artifacts from that past become digital. At the same time, the responsibilities of traditional gatekeepers who have negotiated the ethics of historical data collection and use, such as librarians and archivists, are increasingly being sidelined by the system builders who decide whether and how to provide access to historical digital collections, often without sufficient reflection on the ethical issues at hand. It is our aim to better prepare system builders to grapple with these issues. This paper focuses discussions around one such digital collection from the dawn of the web, asking what sorts of analyses can and should be conducted on archival copies of the GeoCities web hosting platform that dates to 1994.

ACM Reference Format:
Jimmy Lin, Ian Milligan, Douglas Oard, Nick Ruest, and Katie Shilton. 2020. We Could, but Should We? Ethical Considerations for Providing Access to GeoCities and Other Historical Digital Collections. In *2020 Conference on Human Information Interaction and Retrieval (CHIIR '20), March 14–18, 2020, Vancouver, BC, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3343413.3377980

1 INTRODUCTION

In cross-disciplinary collaborations with scholars in the humanities and social sciences, computer scientists are often focused on computational techniques and tools. Given a digital collection of historic interest, they most often lead with the question: What analyses can we conduct to reveal insights? Rarely do we as system builders stop to ponder, *should* such analyses be performed to begin with? However, it is becoming increasingly critical that computer scientists debate this question: Because computational tools are needed to access and study large digital collections, system builders are becoming *de facto* gatekeepers for historical digital material.

This work represents a collaboration between two computer scientists (more specifically, information retrieval researchers), a historian, a librarian, and a social scientist focused on technology ethics, where we grapple with the ethical implications of different types of analyses. We ponder: We could, but should we? The digital collection in question is a web archive of GeoCities, but the issues we explore are applicable to many digital collections that are of historical value. Central to our consideration is the fact that technological tools have advanced tremendously since the creation of the data contained within our archive, enabling analyses that the creators of the content could have never imagined. These analyses may lead to the violation of the information sharing norms expected when the sites were created, and may create new risks or harms for individuals documented in the archive.

To be concrete, we consider three types of analyses, each with their own set of ethical complexities:

- Content-based retrieval, including search based on full-text content as well as other media such as images;
- Large-scale distant reading, including text and link analyses;
- User re-identification, potentially combining analyses of multiple types of content.

The first two types of analyses have already been performed on GeoCities (and other collections) in limited ways, but technologies available today (or that will become commonplace in the near future) can make those analyses potentially more comprehensive and more incisive than ever before. To our knowledge, the third analysis has not yet been attempted, but it is certainly within the capabilities of existing technology.

After setting the stage by providing broader context about both GeoCities and historical scholarship on digital data, we consider the ethics of applying various computational techniques to each of these types of analyses. Researchers struggling with similar questions may reach different conclusions depending upon the ethical framework they choose, intended uses of the data, and the details of the context in which the data were originally created and used. Nevertheless, it is important that they reason about ethics in a systematic and reflective way. We draw upon established frameworks for digital research ethics to illustrate how such systematic reflection might be accomplished.

2 SETTING THE STAGE

The collection in question that triggered our initial inquiry is a web archive of GeoCities. Founded in 1994, GeoCities was a web hosting platform that allowed early users of the World Wide Web to create websites on almost any topic of interest—their love of popular culture (from Buffy the Vampire Slayer fan sites to board games), genealogy, early forms of blogging, and beyond. After its acquisition by Yahoo! in 1999, GeoCities entered a slow death spiral and was ultimately shuttered in 2009. Over its fifteen years of

CHIIR '20, March 14–18, 2020, Vancouver, BC, Canada

ACM ISBN 978-1-4503-6892-6/20/03…$15.00
https://doi.org/10.1145/3343413.3377980

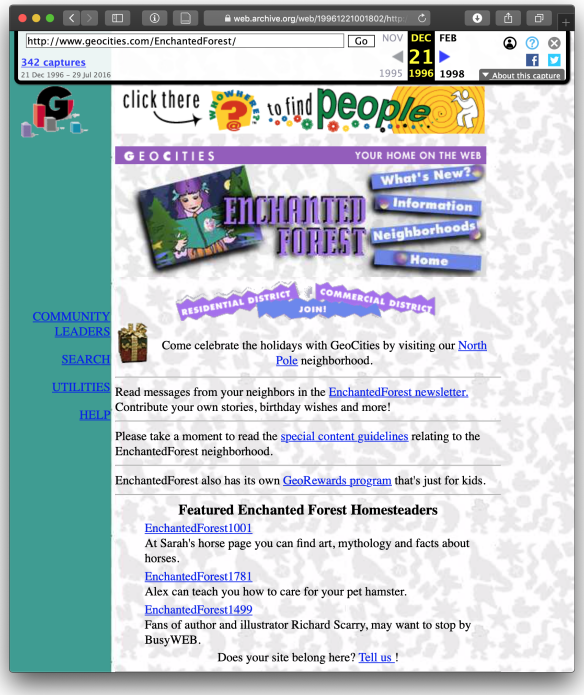

Figure 1: The GeoCities EnchantedForest neighborhood page, from the Internet Archive's Dec. 19, 1996 snapshot.

existence, GeoCities grew to encompass some seven million user accounts, spread over hundreds of millions of HTML pages [33]. It represents a substantial amount of information created by "ordinary users" during the pivotal early years of the web [34]. While early online content creators represented only a minority of the overall population, and are "disproportionately white, male, middle-class and college-educated" [19], so too are traditional archival holdings; these still represent voices that would be largely lost in non-digital memory systems. Today, the GeoCities web archive is available through the Internet Archive's Wayback Machine (see Figure 1), through raw crawls obtained by researcher agreements with the Internet Archive, and as a data dump from Archive Team (a group of "guerilla web archivists").[1]

The structure of GeoCities supports some degree of filtering by date of creation as well as by interest. Pre-Yahoo! GeoCities was arranged into thematic "neighborhoods". When individuals came to the platform to create their page, they were asked to find the right neighborhood for it, such as Athens for academic discussions, the EnchantedForest for pages about or by children (which had enhanced community moderation), Heartland for genealogy, pets, and family pages, or MotorCity for car aficionados. Each neighborhood had 9000 "addresses" (1000–9999); as they filled up, "suburbs" were created to accommodate even more websites. URLs would thus look like `geocities.com/EnchantedForest/Grove/3891`. After its 1999 purchase, Yahoo! discontinued the neighborhood structure and shifted GeoCities towards a "vanity" URL structure (e.g., `geocities.com/mysite`).

[1] Available at https://archive.org/details/archiveteam-geocities.

In 2015, we came into possession of a complete web crawl of GeoCities from the Internet Archive, dating to 2009, just before the site closed. While many might be familiar with the Internet Archive's Wayback Machine, which provides full-text search only for very limited content, we reached a research agreement with the Internet Archive to use the raw Web ARChive (WARC) files, the ISO standard container format for web archives. Totaling approximately 4TB, these files contain all the information needed to roughly reconstruct GeoCities as it was at the time of the web crawl: not only the HTML pages, but images and other multimedia content as well. Thus, the types of analyses that we are able to conduct are far more comprehensive than what a typical scholar might manage with only the publicly-available Wayback machine.

3 THE HISTORICAL PERSPECTIVE

The process of creating historical knowledge is complicated. Events happen—for example, by the time you are reading this sentence our act of writing it is in the past—but in general, unless past events are documented and preserved, they will not become part of history. The default condition of events, then, is that they happen and then are eventually forgotten. In other words, historians generally operate in a context of source *scarcity*. We wish we had more information about the past, but we do not, because libraries and archives are very limited in the physical collections they can accession, preserve, and provide access to.

The sheer scale of web archives such as GeoCities means that our historical record is undergoing a dramatic transformation, a theme explored in recent monographs [6, 34]. While web archives are not a magic bullet—the majority of things that are happening are still not preserved, social media or websites aside—they do represent a marked increase in the amount of information being preserved. This is especially important for historians studying society and culture. Coupled with the dramatic increase of digitized material, historians are increasingly needing to develop new skills to deal with this deluge of digital information. The historical profession has transformed from one that suffered from information scarcity, to one of abundance [16, 42, 43].

The scale and breadth of GeoCities raises particularly significant historical opportunities. One of the co-authors is a historian of childhood and youth by training. While children are not traditionally well represented in the historical record—children do not generally leave sources, as diaries are rare, and subsequent oral histories are filtered through the lens of adulthood—GeoCities alone offers tens of thousands of web pages in the EnchantedForest neighborhood, "a place for and by kids" (see Figure 2). Furthermore, most GeoCities sites were created by "everyday" people, with the provisos noted in the previous section. Social historians are concerned with history from the "bottom up", to emphasize the lives of individuals. GeoCities thus presents important opportunities to understand society and culture in the late 1990s.

As many historians principally work with curated materials, their engagement with formal research ethics has been relatively limited. Consulting material in a formal archive or library typically does not require ethics review because archivists and librarians will generally have navigated the maze of donor agreements and restricted materials, meaning that historians can largely be confident

in their right to access and explore material. While this does not mean that historians are immune from ethical questions—consider, for example, discussions among historians about ethical responsibility to the dead [36]—it does mean that by and large historians are suddenly without their traditional supports upon entering the web world. For one, web archives typically do not have donor agreements and few broad definitions of restricted materials. Even more importantly, web archives are often "archives" without archivists; what we are really talking about is a web corpus. It is we, as system builders, who stand in the place of the archivist—and that is a role for which many of us might feel more than a tad under-prepared.

4 ETHICAL FRAMEWORKS

A first step for scholars and system builders is to choose one or more ethical frameworks to guide decisions about how (and whether) to search and analyze web data. The choice of ethical frameworks should be guided by a researcher's own context as well as the cultural fit of the framework to the research.[2] As American and Canadian researchers with backgrounds in history and technology ethics, we have chosen to consider guiding frameworks that dominate US and Canadian research ethics.

Perhaps the closest parallel framework for historians are guidelines developed for oral history. As scholars meet with people in formal settings to discuss their past experiences, they draw on a large and growing body of support to guide them in their practices, emphasizing care, the right of withdrawal, and the cultivation of meaningful community relationships [18]. Digital historians have explored the impact of making material accessible online, both through panel discussions and scholarship as well as through tool development [28]. The Mukurtu Content Management System, in particular, is designed around the cultural frameworks of indigenous peoples. It allows people to "define the terms of access to and distribution of their cultural materials… [f]or example, a piece of content uploaded by an individual may be designated for women only. Or, an image of a male initiation ceremony returned from a national museum may be eligible for viewing by elder men only." Chief among its values is being open to "constant negotiation" [7]. Scholars who use Twitter data have also been exploring ethical challenges inherent in the collection, and more importantly, publication of social media data; a recent ethics consultation drew no firm rules but laid out a number of unresolved and critical questions [49].

In our American and Canadian research context, another logical place to seek guidance is the framework provided by the Belmont Report [39], and updates to it for Internet research by the Association of Internet Researchers (AOIR) [27]. We also illustrate how contextual integrity [37], an ethical framework developed to respond to privacy concerns in digital data, might further support inquiry into the ethics of various analyses.

4.1 The Belmont Report

In the US and Canada, questions of how to ethically conduct research using data from and about people is guided by the Belmont Report, a set of ethical guidelines for biomedical and behavioral research developed in response to research ethics scandals in the mid-20th century. Belmont emphasizes three principles: respect for persons, beneficence, and justice. However, the ways that Belmont has been codified into law and institutional practice in the US and Canada reflect assumptions drawn from laboratory experiments and biomedical research [30]. While the report itself does not define what sorts of research are and aren't covered, critically for our study, the US Common Rule, which governs research ethics at any US research institution that receives public money, does. First, the Common Rule [39] defines "human subjects" as

> a living individual about whom an investigator (whether professional or student) conducting research: (i) Obtains information or biospecimens through intervention or interaction with the individual, and uses, studies, or analyzes the information or biospecimens; or (ii) Obtains, uses, studies, analyzes, or generates identifiable private information or identifiable biospecimens.

Already this definition presents problems for GeoCities. Many of the individuals in the collection are not be identifiable, meaning they are likely not seen as "human subjects" as defined by the Common Rule. For those that are identifiable, do GeoCities pages constitute "private" information? The Common Rule also provides guidance here, defining private information as

> information about behavior that occurs in a context in which an individual can reasonably expect that no observation or recording is taking place, and information that has been provided for specific purposes by an individual and that the individual can reasonably expect will not be made public…

Furthermore, the Common Rule declares exempt from regulation "secondary research uses" of existing data if "identifiable private information or identifiable biospecimens are publicly available". GeoCities was, and still is, publicly available, so research on this data constitutes a secondary research use.

In Canada, research ethics are informed by the Tri-Council Policy Statement on Ethical Conduct for Research Involving Humans. It declares that board review "is not required for research that relies exclusively on cyber-material, such as documents, records, performances, online archival materials, or published third party interviews to which the public is given uncontrolled access on the Internet and for which there is no expectation of privacy", noting that researchers should consult terms of service and that sites with "restricted membership" mean that there is a higher privacy expectation [40].

A US Institutional Review Board or Canadian Research Ethics Board would almost certainly give researchers consent to, for example, build a search engine for the GeoCities websites and conduct research on the collection. However, this isn't the end of the story.

4.2 The AOIR Guidelines

Traditional research ethics as guided by the Belmont Report has been criticized as being badly adapted to digital information. The Common Rule's definition of public information doesn't take into account the changing social information norms around websites and blogs, social media profiles and posts, and other pervasive data trails [30, 48]. There is increasing empirical evidence that users of, for example, social media sites are surprised and displeased to find their public contributions and actions subject to observational [14] and intervention [17] research.

[2] For a primer on intercultural frameworks, see Ess [13].

In response to early recognition of the ethical challenges of Internet research, an AOIR working group developed guidelines in 2002. The guidelines received a major update in 2012, and a recent update in 2019. The AOIR principles focus on context and declare that researchers must:

- Weigh potential harms to research communities or subjects according to their specific context;
- Balance the rights of subjects with the social benefit of the research;
- Become more cautious as the vulnerability of the community being studied increases.

Among the issues that the AOIR guidelines ask researchers to consider are the ways that context, social vulnerability, methods of access to data, analyses, and potential findings create harms and risks. They also ask researchers to reflect on how they recognize the autonomy of others and acknowledge the equal worth and dignity of the research subjects.

4.3 Contextual Integrity

AOIR's guidelines offer perhaps the most comprehensive currently available basis for grappling with questions of whether and how to provide new access to, or analyses of, Internet data. However, the AOIR guidelines purposefully do not prescribe a set of rules to follow, because challenges such as defining the context of the research, deciding where Internet data falls on the continuum of public to private, recognizing vulnerable populations, and deciding how best to respect the worth of research subjects while balancing the social good of the research are complex tasks.

One framework that can provide helpful empirical direction for answering some (but not all) of these hard questions is contextual integrity. Contextual integrity is an approach specifically adapted for considering the ethics of digital data sharing [37, 51]. It argues that people have expectations for how their data will be used and shared in particular social contexts, and that violations of those expectations should be carefully considered. In particular, violating established information norms is warranted only when it helps to meet the goals of the social setting in which the data was created. For example, data sharing between a doctor's office and an emergency room team may surprise an individual who expected their medical records to be kept in confidence, but it would be reasonable if it furthered the goal of patient health.

In an article applying contextual integrity to digital research ethics, Zimmer lays out a series of questions researchers must ask themselves [52]. These focus on describing the information flows of a research project, including the data's information subjects, senders, and recipients. It also involves examining the social context in which the information was created. If new research represents a violation of expected norms, then researchers must consider:

> the moral and political factors affected by the new practice. How might there be harms or threats to personal freedom or autonomy? Are there impacts on power structures, fairness, justice, or democracy? In some cases, the results might overwhelmingly favor accepting or rejecting the new practice, while in more controversial or difficult cases, further evaluation might be necessary… How does the new practice directly impinge on values, goals, and ends of the particular context? If there are harms or threats to freedom or autonomy, or fairness, justice, or democracy, what do these threats mean in relation to this context?

Even contextual integrity, however, does not guide our reflections completely. In particular, the effects of time on information norms and expectations are under-theorized. How should the affordances of nearly three decades of technological progress change our accounting of user expectations, information flows, and values and goals of the context?

4.4 Guiding Principles

In previous work [31, 34], we have melded oral history and AOIR guidelines [26, 27] into two guiding principles. These are:

- **Context**: What was the context in which content was created? What data uses might the creator have expected? In some cases, this can be inferred from a close reading of the text itself, or from the hyperlink structure of a website. For example, in GeoCities, a page that was linked to by hundreds of other GeoCities webpages might not have a reasonable expectation of privacy; a webpage that was largely undiscoverable then, and is only discoverable due to modern search technology, should be handled with considerably more care.
- **Scale**: What is the scale of the research being conducted? Conducting distant reading research using techniques such as large-scale entity extraction or link analysis across thousands of webpages [8] is not without ethical considerations, but these analyses are very different than a close reading or providing detailed, attributed quotations to a person's individual homepage.

Our GeoCities project, as well as reflections on justice encouraged by AOIR guidelines, suggests an additional guiding principle: historical representation, or the value of an inclusive and diverse historical record. While the web, particularly the web of the late 1990s, is not a magical, all-encompassing place (there are, of course, serious questions to be asked about who is represented on the web and who is not, who accesses the web and who does not, and various ways in which voices are amplified and suppressed), it still does allow for the historical inclusion of some everyday people in dramatic fashion. We cannot simply abandon studies of social media platforms or websites like GeoCities due to privacy concerns, as that will have the effect of making the historical record skew towards the powerful: corporations, governments, elites, those who have established recordkeeping programs and digital preservation plans, or can actively shape their memory and legacy moving forward. The voices of the LGBT community in the late 1990s web or children writing about their experiences in public school, for example—these are voices that are critical to capturing the diverse and inspiring human experience.

5 CONTENT-BASED RETRIEVAL

The American pundit Yogi Berra is credited with observing that "In theory there is no difference between theory and practice. In practice there is." If we are to apply these theoretical perspectives to practical applications, we must consider the details of those applications. In this section we start by considering issues related to providing content-based retrieval capabilities for GeoCities data.

We broadly consider the case in which a scholar has a particular information need and wishes to interrogate the collection to retrieve relevant content. Critical to the scoping of this discussion is that a scholar is involved, which sets an upper bound on the amount of material that can be "consumed". We also leave out-of-scope more fraught cases such as illegal uses of data (e.g., to steal identities or increase accuracy at guessing online security questions). In the most common case, the scholar's information need can be expressed as a keyword query to retrieve relevant web pages (full-text search), but we also consider other possibilities such as image search, which can exploit technologies that would have been close to "science fiction" from the perspective of late 1990s users.

Since the demise of GeoCities itself, there has been a searchable GeoCities archive that went live by 2010, and a second GeoCities archive that is currently indexed by major search engines such as Google and Bing. So, despite the issues we raise here, the fact is that GeoCities web pages have been searchable for most of the past 25 years. However, questions of ethics are about what we ourselves should do—and thus the existence of these resources does not absolve us from the need for reflection.

Notably, GeoCities is not the only historical collection for which researchers have grappled with these questions. Some of the same issues that we raise here also entered the public discourse back when previously ephemeral USENET news posts first became permanently searchable [47]. For example, the Deja News Research Service, a fairly comprehensive searchable archive of USENET that was first introduced in 1995, remains available today (as part of Google Groups), although there have been some issues with continuity of search capabilities in the face of changing technology [5]. If we do elect to create research services to search GeoCities, we might do well to be humble and to define the period over which we intend to provide that capability.

Applying principles of context, scale, and historical representation to the case of content-based search yields a conclusion that many—but not all—such uses would be ethically appropriate. Central to our consideration of context is the *information flow norms* for GeoCities users during its heyday. Could they have understood that their content could have been easily found by a search engine in response to queries by *anyone*? Answering this question breaks down into two components: the technical limits of search capabilities of the time and the contemporaneous user perceptions. As we discuss, neither set of details is quite clear, making the context harder to reason about than we might wish. Scale and historical representation are more straightforward.

5.1 Full-Text Search: Technical Capabilities

Full-text search on the web has existed since the earliest days in the history of GeoCities. By 1994, when GeoCities launched, search engines included Infoseek, Webcrawler, and Lycos; these would be joined in 1995 by Altavista and in 1998 by Google. In addition, GeoCities provided a search function that is still visible—if not functional—from a 1996 Internet Archive crawl.[3] While it is hard to know exactly how expansive or capable some of these search engines were, results from a February–November 1996 study do shed some light on their capabilities—and their dramatic growth. In one example, a search for the keyword "embargo" on Lycos in February 1996 found 40 hits and by November over 62 thousand [41]. Indeed, crawler coverage varied dramatically by search engine: in 1995/1996, this ranged from 1.5 million pages (Excite) to almost 20 million (Lycos) [9]. Search engines were becoming increasingly powerful throughout this period, but they were nowhere near as capable then as they are now. Lending credence to this, as late as 1999, we see books in the popular press informing GeoCities users that while search engines would be indexing their web pages, they should engage in various search engine optimization techniques to help improve their rankings within those results [45].

5.2 Full-Text Search: User Perceptions

With some of the earliest GeoCities sites having been created in an era when web search was in its infancy, some early users may have provided personal information in a public environment but relied on "privacy by obscurity" (e.g., by relying on their sites to be hard to discover outside of closed peer groups connected by webrings or hyperlinks). Indeed, the existence of technologies such as webrings—which have all but disappeared today due in part to better search capabilities—suggests that there were serious contemporary limitations around discoverability. Moreover, the reality was that web penetration rates were much lower prior to GeoCities' peak, so even the concept of "public" was not nearly as encompassing as it is on today's web. For example, one study found that individuals within the LGBT community provided intimate details of their personal life—despite not being out within their offline communities at the time [29]. Thus, there is at least some evidence that a "typical" user in the mid to late 1990s may not have realized that their GeoCities content was easily findable.

However, we find similarly anecdotal evidence suggesting that at least some users from that period were fully aware of the reach of web search engines. According to Wikipedia, the notion of "egosurfing"—searching for oneself on the web, perhaps better known as "vanity searching" today—was coined by Sean Carton in 1995 and first appeared in print as an entry in Gareth Branwyn's March 1995 Jargon Watch column in Wired.[4] This demonstrates that, since the initial days of the web, there have been at least some users who were not only aware of the extent of search engine coverage, but wholly embraced it. In a 2000 article, Dent [11] remarks that the term first appeared in the Oxford English Dictionary in 1998, which suggests that by that time, a sizeable portion of Internet users would have been familiar with the concept.

There are considerations, however, beyond user perceptions. Just as importantly, GeoCities sites were created by a wide range of people including minors, who gave no explicit consent for their pages to be studied, putting the onus on historians and other scholars to carry out an ethical assessment themselves. In addition, we might reasonably expect users who created their sites in these early years of GeoCities to be unaware that their sites have been *preserved*.

5.3 Beyond Full-Text Search

While image search capabilities have existed since the early days of the web—the ability to type in a few query terms and retrieve relevant images—early systems for the most part exploited textual

[3] https://web.archive.org/web/19961221013515/http://www.geocities.com/search/

[4] https://en.wikipedia.org/wiki/Egosurfing

features such as anchor text, metadata tags, file names, and image captions [15]. Thus, there wasn't likely much that could be discovered with early image search engines beyond what would already be discoverable with full-text search.

Image search, however, improved over time. In addition to textual features, systems began to incorporate features that were directly extracted from image content, for example, colors, shapes, and textures. So-called content-based image retrieval was initially developed in the 1980s, but scalability had always been an impediment to large-scale deployments. Image analysis is much more computationally-intensive than text analysis, but the relentless pace of hardware advances over the years made content analysis increasingly practical. The early 2000s marked the arrival of a new capability known as query-by-example: the user could submit an image as a query to retrieve "similar images". Here, we have an example of an entirely new capability that did not exist during the heyday of GeoCities, and most GeoCities users surely would not have anticipated that images on their pages could be investigated in this manner. Such capabilities have been a boon to scholars, for example, studying the spread of early Internet memes.

The resurgence of neural networks and deep learning since the early 2010s have led to no less than a revolution in many areas in computer science, including computer vision and natural language processing [24]. A poster child for the prowess of deep neural networks is the dramatic increase in the accuracy of object recognition from images. Features from the neural networks that perform these tasks can be extracted and indexed as the basis of search [1]. Image search capabilities have expanded by leaps and bounds as a result. Not only can we, for example, find pictures of dogs with great accuracy, we can restrict searches to specific breeds, in specific settings (e.g., on a beach), and even engaged in certain activities (e.g., chasing a ball). Not only can we find images of people, but we can retrieve images of *a specific person*. These capabilities would surely not have been anticipated by a typical mid-1990s web user. It is unlikely that users who posted pictures from their adolescence (engaged, for example, in socially questionable behaviors) would expect a search engine to find those pictures a quarter century later.

5.4 Ethical Considerations

Applying a contextual integrity analysis to content search highlights that today's search tools likely break some of the information flow norms on the 1990s web. Users may have expected some of their text to be discoverable by search engines, but almost certainly could not have expected the accuracy and granularity of today's search engines, or content search by other means, such as the advanced image search capabilities discussed above. But as Zimmer [51] and Nissenbaum [37] point out, just because an action breaks contextual integrity does not make it unethical. The next question is to determine whether the goals of the research are consonant with the values of the original context. In many cases, improving search arguably supports the goals and values of the GeoCities context. GeoCities was one of the first widely-used Internet publishing platforms, and its goals were sharing and communication. One of the key planks of GeoCities' marketing was the sheer number of users and visitors who could visit a new user's homepage. The founder of GeoCities, David Bohnett, would later recall that "we all have something to share with each other, which enriches both their lives and ours as well" [38]. For example, in 1996 alone, GeoCities noted that their "incredibly high volume of traffic assures the highest possible visibility for your home page"[5] and the rhetoric behind the site was to "link people and their ideas together in a way that was never possible before."[6] This was not hyperbole—by the middle of 1998, GeoCities was one of the top ten most popular sites on the web; and by 1999, it was the third most popular site [32].

GeoCities' emphasis on discoverability can be seen beyond metrics, in the neighborhood structure of the site itself. Users were never supposed to be alone, as the platform was both marketed and designed to function as a virtual community. Each set of addresses was assigned a volunteer community leader who was supposed to reach out to new "homesteaders", ask them if they need any help, and to visit their pages and offer suggestions for improvement. In short, GeoCities marketed itself as the place to be if you wanted to make a homepage that was discoverable. Because sharing and communication were part of the original intent of the GeoCities context, we argue that contextual integrity supports building a search engine for this material. In addition, considerations of historical representativeness argue for enabling search access.

However, considering our principle of scope complicates the ethics of using a search engine to provide granular, specific access to GeoCities pages. Using a powerful search engine, researchers could study individuals, making GeoCities participants research subjects in ways they might never have anticipated. To address these concerns, a search tool could be sensitive to contextual integrity in various ways. Perhaps granular search capabilities could be made most widely available for GeoCities sites that originally had the greatest number of inlinks, and thus would have been the most discoverable in their original context. Less "popular" GeoCities sites might also be rendered broadly searchable if the search engine could be made sensitive to particular vulnerabilities, e.g., webpages created by children or by members of a marginalized group. Researchers who agree to anonymize their results might be given further access to the most comprehensive search settings. Finally, researchers who use a search tool on GeoCities data should consider the relative vulnerabilities of the people or groups they are studying, and subsequently how they share and publish the material they find. Researchers should consider whether the passage of time has mitigated (or exacerbated) particular vulnerabilities. For example, users who were once children are now adults. These adults may no longer be embarrassed by childish crushes, but they might be more negatively impacted by revelations of illegal activity. Researchers should consider obscuring identifying details of individuals, particularly individuals from marginalized communities, or whose behaviors might be socially stigmatized.

6 DISTANT READING

Literary scholar Franco Moretti in 2000 proposed the idea of "distant reading". Whereas scholars traditionally engage in "close reading", or readings of individual novels, he instead proposed trying to understand much larger systems through emerging computational

[5] https://web.archive.org/web/19961219234328/http://www.geocities.com/BHI/about.html

[6] https://web.archive.org/web/19961221005714/http://www.geocities.com/homestead/FAQ/faqpage1.html

methods. As he argued, a large field can not be "understood by stitching together separate bits of knowledge about individual cases, because it isn't a sum of individual cases: it's a collective system, that should be grasped as such, as a whole" [35]. The digital humanities, an umbrella term for scholars engaged in the use of new and emerging technologies for the creation and dissemination of knowledge in the humanities, has found wide application for many natural language processing and information retrieval techniques. Examples include distant readings of transnational commodity flows [23], parliamentary proceedings [4], musical recordings [46], and hundreds of thousands of court transcripts [10].

In our context, distant reading techniques are wide ranging. Examples include finding the frequency of words that appear in a collection of documents and visualizing them as a word cloud, using Latent Dirichlet Allocation (LDA) to identify topics within texts [16], and prosopography [21] based on finding references to entities, events, and relationships using information extraction techniques. In web collections that lend themselves to network analysis, a range of techniques can be employed to identify the documents that are the most central or connected. Finally, image analysis—which, as we have discussed above, has become increasingly sophisticated—allows scholars to conduct analyses at scale, for example, to find objects of research interest [50].

For GeoCities, historical scholarship is already beginning to appear: explorations of community in the context of the "virtual community" debates of the 1990s [32] and of public/private boundaries in diaries [2], for example. Other potential questions include studies of 1990s popular culture, how politics and elections were filtered through websites, and more generally, exploring how people made sense of the new medium of the World Wide Web. Thanks to recent advances in scalable, easy-to-use computational tools [25, 44], we are able to apply distant reading techniques to GeoCities.

With nearly 200 million interconnected HTML pages, one obvious starting point is network analysis. By extracting and then visualizing hyperlinks within the collection, we can begin to reveal internal community dynamics. Figure 2 shows such an analysis applied to the EnchantedForest neighborhood, created using the Archives Unleashed Toolkit [44] and the open-source Gephi network visualization platform [3], where website labels are sized according to PageRank. Thanks to readily available tools and tutorials, humanities scholars with limited technical training are able to conduct such analyses themselves. From the visualization, a number of sites immediately jump out. The largest ones, such as EnchantedForest/Glade/3891, are linked from large swaths of the community; these represent volunteer hubs, instructional pages, frequent guestbook commentators, and those centrally located within webrings or reciprocal networks. Conversely, we also find relatively private sites: those that had no incoming links from within GeoCities. One approach that we have taken is to identify the top 100 sites in each neighborhood in terms of PageRank and then manually examine them. This combination of distant and close reading is effective in revealing major themes in those communities.

Text analysis also presents interesting possibilities when applied to GeoCities content. For example, we could exploit topic modeling to determine what topics are prevalent within a neighborhood or group of websites. In the TelevisionCity neighborhood, we find a topic that includes the words "Rachel Chandler Ross Monica Phoebe

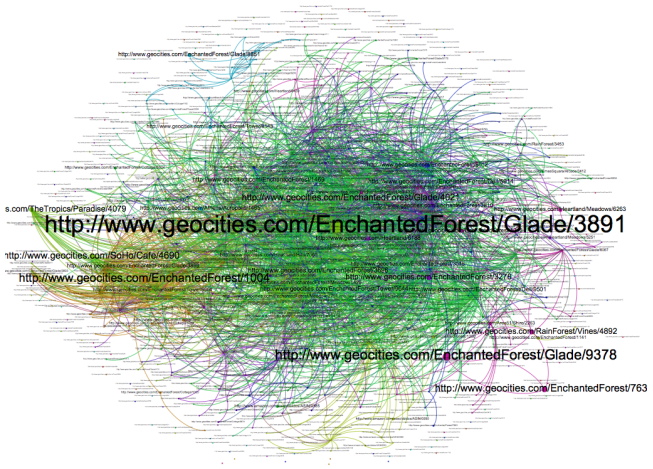

Figure 2: Visualization of the network structure within the GeoCities EnchantedForest neighborhood.

Joey": the central characters of the NBC sitcom *Friends*. We can then begin to manually explore pages that discuss this show, either through close reading or by creating a sub-collection to explore. Similarly, entity analysis could allow us to find websites discussing a given location, person, or organization. Such techniques are a boon for cultural historians.

6.1 Ethical Considerations

At first blush, distant reading seems to mitigate some of the ethical concerns raised by search. From Section 4.4, our principle of scope, paired with reflections on power and justice, would argue that many types of distant reading do not inspire the same ethical concerns as search intended for close reading. Studying *structures* rather than *people* produces benefits for society without subjecting individuals to the kind of scrutiny that raises contextual integrity and human subjects concerns. Yet, in carrying out this research, we can also begin to lose the context of the topic being studied or even the pages that are inevitably found.

Compounding the challenge of a loss of context is the fact that there is rarely a firm line between distant and close reading. Just as search ends up with pages that are closely examined by scholars, so too do most forms of distant reading. Even if the patterns found are "at scale" without revealing individual identities, at the publication stage, practices in historical research generally require citation to individual pages as evidence. It is important that readers be able to visit a primary document, at least in theory, to form their own interpretation of the narrative being advanced.

A few examples can help illustrate this point. When a topic model identifies exemplars for a particular topic, there is a danger that we might publicize those pages as such without a firm contextual understanding of where they fit in the broader collection. Imagine a topic model that has found a sensitive topic (e.g., self-harm) and representative pages. When this topic is then discussed in a paper, would it be appropriate to cite these pages—and thereby the individuals who created them—as exemplars of a much-larger pattern within a web archive? Particularly when the scale prohibits a researcher from reading every single page within said topic, or even

knowing what features of the page had resulted in model's label assignment? Labeling a page as an exemplar of self-harm might feel like an accusation to someone who did not identify with that term. Similarly, PageRank might reveal popular websites about a sensitive topic, but again, the scale of the analysis may prohibit researchers from gaining a nuanced sense of why the PageRank is so high, or the context of the incoming links.

Researchers performing distant reading should therefore take similar precautions as those performing content search for close readings. Do the goals of the research mesh with the values of the GeoCities community being studied? Even distant reading of sensitive topics should reflect on the relative vulnerability of the people behind the pages, as encouraged by AOIR guidelines. Exemplar pages should be anonymized to the extent possible.

A final, particularly fraught, area of distant reading comes when it is paired with *re-identification* (more below). What if this sort of analysis—a topic model, as noted above—revealed a *named* individual's site to be at the center of sites interested in illegal drug use? Or in promoting political dissent?

7 USER RE-IDENTIFICATION

All of the analyses and technical capabilities discussed above come together in the use case of re-identification. By this, we mean computationally establishing correspondences between pages in GeoCities and persons today, either alive or deceased, in a *personally-identifiable way*. The evidence for this correspondence may range from circumstantial, for example, photographs that seem to show the same person, to indisputable, for example, matching social security numbers.

We see a variety of ways in which user re-identification can be deployed. We might envision a "re-identification search engine" to which a user submits a request to find a specific person in the GeoCities dataset. The "query" might include some personally identifiable information of the person in question, e.g., email addresses, social media handles, photographs. Such a search engine might be used in at least three ways: users could look for themselves, users could look for someone else, or users could attempt to match as many GeoCities pages as possible with current identities.

In the first case, there are technical means to establish someone's identity and acquire informed consent. For example, users can explicitly give an app permission to gather their personal data from Facebook.[7] In fact, one of the most frequent requests that we hear when talking about GeoCities data is finding the abandoned pages that the person asking created long ago.

In the second case, a re-identification search engine could also be used to search for other people. A common use case would be genealogical research. In another plausible use case, researchers might gain important insights into the early roots of violent radicalization by studying the childhood websites of known terrorists. Or to take a page from recent news in the US and Canada, another use might be finding insensitive or inappropriate imagery posted by or of now-powerful politicians.

In the third case, we can imagine a tool that tries to exhaustively match all persons in the GeoCities dataset with more current publicly-available data (e.g., web crawls of personal homepages, social media profiles, etc.). The output of this massive data mining operation would be a database of correspondences between known individuals (represented by, for example, a Facebook handle) with pages in GeoCities. Such data could then be made available in our hypothetical re-identification search engine, or the raw data might even be made available for public download. This is reminiscent of the 2015 Ashley Madison data breach,[8] when a large data dump was leaked containing identities of individuals who had signed up for the commercial website to ostensibly engage in extramarital affairs. Subsequently, websites emerged to allow jilted spouses to search the data dump.

We are confident that technologies exist to perform re-identification on a massive scale today, even if no one (to our knowledge) has yet attempted to realize the scenarios that we have sketched above. Obvious starting points include using existing unique identifiers such as email addresses and social media handles. This translates into a massive $N \times M$ string matching problem, which is well within the capabilities of modern hardware, even on terabytes of data (or more), particularly if we exploit techniques such as sketches and similarity blocking to reduce the quadratic search space. Note that re-identification may be accomplished transitively rather than directly, for example linking GeoCities not directly to Facebook, but via digital traces on Friendster or other now-defunct sites with available web archives.

Beyond text, other features can be brought to bear to aid in re-identification. As with other types of social networks, the community structure revealed by the link graph—both historical link graphs and the modern one—can be a particularly powerful way of resolving ambiguous cases. Facial recognition technology, from the same revolution in deep learning that powers the image search capabilities discussed in Section 5.3, would also be useful. Advances in natural language processing that consider stylistic features (for example, in an authorship attribution task) can help in cases where there is a sufficient volume of text. By combining multiple, heterogeneous features (text, network, image, etc.), computational models today can potentially make inferences that would not previously have been possible—and certainly beyond the expected norms of content creators in the late 1990s.

A final point on re-identification technology: it bears emphasizing that the output of any computational model can only be considered *hypotheses*, albeit with varying confidence based on the strength of evidence. The "ground truth judgment" of whether any two identities are indeed the same resides solely with the person in question, for example, politicians confessing that it was indeed them who engaged in the questionable behavior depicted in purported photographic evidence. This is an important point because computational models inevitably make mistakes—for example, email addresses and social media handles can be abandoned and then reclaimed by unrelated individuals, creating connections that have no basis in reality. Here, the Ashley Madison data breach is potentially instructive, as people with similar-sounding names and similar emails have proved easy to confuse, and there are known cases of users who had accounts created without their consent. These mixups can have grave consequences. The imprecise nature

[7] There is, of course, the separate issue of what such a search engine can and cannot do with personal data once acquired.

[8] https://en.wikipedia.org/wiki/Ashley_Madison_data_breach

of computational models means that there will inevitably be cases of mistaken attribution, with all the attendant risks discussed above.

7.1 Ethical Considerations

As we have argued above, there are potentially valid reasons to work on re-identification. We could, but should we?

Re-identification of a page posted under an alias is clearly a violation of contextual integrity. Indeed, this is just the sort of scenario that the so-called "right to be forgotten" enshrined in the European Union's General Data Protection Regulation (GPDR) seeks to manage. Indeed, the right as defined in Europe is not truly a right to be forgotten but rather a right not to be re-found. The question then becomes whether the violation of contextual integrity is warranted by supporting either the goals of the original social context, or the potential social good of the research. Revealing formerly anonymous participants does not support some of what we might reasonably expect were the original norms or goals of most GeoCities users. Looking at the intersection of digital forensics (an applied field around the investigation and recovery of information and materials in digital devices) and archives, a 2010 report [22] commissioned by the Council on Library and Information Resources, which examined digital forensics and its applications to archivists and curators, devoted a large section of the report to ethics:

> Another aspect of social networking that must be considered in relation to concerns over privacy relates to using aliases. Since the inception of the Internet, many individuals have preferred to use aliases to protect their identity. This practice could cause problems for archivists, digital curators, and researchers interested in identifying the various online presences, such as blog postings or a Facebook page, of individuals whose papers are in a repository's collection.

In cases where researchers have explicit permission of an individual to re-identify their old site, the ethical issues are minimal. Such permission can function as informed consent, and re-identifying someone's page at their own request is therefore reasonable.

Re-identification of the pages of others, however, should be undertaken with much more care, and with the knowledge that it violates user expectations and contextual integrity. In cases where a larger social good is served by re-identification, there may be reasons to do so. A biographer of a historically-significant individual, for example, could make a compelling case that the value to our historical knowledge might be beneficial. Should a political leader, for example, enjoy privacy protections over their early web presence (provided they were an adult at the time)? In Canadian historiography, for example, access to the personal diaries of Mackenzie King—Prime Minister during the Second World War—has considerably enhanced our understanding of both his leadership and the broader political, social, and cultural context in which he operated. While there was a "an ambiguous demand in Mackenzie King's will to have the Diaries destroyed",[9] ultimately librarians, archivists, and historians decided that the public good was better served by preserving them for research. This could perhaps provide a parallel to the questions before us today.

[9]See http://www.bac-lac.gc.ca/eng/discover/politics-government/prime-ministers/william-lyon-mackenzie-king/Pages/diaries-william-lyon-mackenzie-king.aspx and also Dummitt [12].

It is not just the records of the politically powerful, of course, who we need to consider. Our guiding principle of historical representation, for example, suggests that researchers may need to re-identify a large subset of GeoCities pages to study a particular group who would not otherwise be reachable. In such cases, researchers must reflect upon, and write about, why such re-identification is justified, and what steps they have taken to respect the (formerly anonymous) context of information creation. Researchers might consider consulting members of the community they are studying [20], for example, to be sure to document and mitigate potential re-identification harms. Furthermore, researchers should almost certainly not make a re-identified data dump publicly available without access controls.

8 CONCLUSION

We have proposed three ways of thinking about some of the ethical choices inherent in the systems that we can now build. The Belmont Report's principles, together with the Common Rule's definitions, provide a set of guidelines for many research settings, but that perspective is insufficient to address the full range of challenges posed by modern technologies for manipulating large-scale historical digital collections. The Association of Internet Researchers has offered additional guidelines that elucidate the range of challenges. Paired with contextual integrity, which provides a way to account for the expectations of creators, these guidelines provide useful tools for reasoning about ethics.

We have selected three specific settings to illustrate the application of contextual integrity to scholarly research: content-based search, distant reading, and user re-identification. While we have articulated some preliminary conclusions and our rationale for arriving at those conclusions, our analysis is necessarily incomplete, since ultimately it is the system builder and the scholar who must make these choices together for their specific application. However, our goal has not been to prescribe, but rather to illuminate the range of factors that should be considered, and how contextual integrity can provide a useful framework to guide these choices in a principled and reflective manner. We see this as the start and not the end of a discussion.

The time for these conversations is now. This paper has demonstrated that working with web archives raises some less well studied dimensions of research ethics. Whether a researcher is engaged in implementing search, distant reading, or user re-identification, working with—and publishing on—these materials can raise many fraught questions. Even a researcher with the full approval of a (20th-century) ethics review process can suddenly find themselves at the center of an ethics maelstrom. We hope this paper might help researchers identify, reflect on, and proactively address some of these challenges.

9 ACKNOWLEDGMENTS

This research was supported by the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, the US National Science Foundation (grants 1618695 and 1704369), the Andrew W. Mellon Foundation, Start Smart Labs, and Compute Canada. Opinions expressed are the authors' and do not necessarily reflect those of the sponsors.

REFERENCES

[1] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky. 2014. Neural Codes for Image Retrieval. In *ECCV*. 584–599.

[2] J. Baker. 2020. GeoCities and Diaries on the Early Web. In *The Diary*, B. Ben-Amos and D. Ben-Amos (Eds.). Indiana University Press.

[3] M. Bastian, S. Heymann, and M. Jacomy. 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks. In *ICWSM*.

[4] K. Beelen, T. Thijm, C. Cochrane, K. Halvemaan, G. Hirst, M. Kimmins, S. Lijbrink, M. Marx, N. Naderi, L. Rheault, R. Polyanovsky, and T. Whyte. 2017. Digitization of the Canadian Parliamentary Debates. *Canadian Journal of Political Science* 50, 3 (2017), 849 – 864.

[5] M. Braga. 2015. Google, a Search Company, Has Made Its Internet Archive Impossible to Search. *Motherboard* (Feb. 2015). https://www.vice.com/en_us/article/jp5a77/google-a-search-company-has-made-its-internet-archive-impossible-to-search

[6] N. Brügger. 2018. *The Archived Web. Doing History in the Digital Age.* MIT Press.

[7] K. Christen. 2011. Opening Archives: Respectful Repatriation. *The American Archivist* 74, 1 (2011), 185–210.

[8] H. Christenson. 2010. HathiTrust: A Research Library at Web Scale. *Library Resources and Technical Services* 55, 2 (2010), 93–102.

[9] H. Chu and M. Rosenthal. 1996. Search Engines for the World Wide Web: A Comparative Study and Evaluation Methodology. In *ASIS*. 127–135.

[10] D. Cohen, F. Gibbs, T. Hitchcock, G. Rockwell, J. Sander, R. Shoemaker, S. Sinclair, W. Turkel, C. Briquet, J. McLaughlin, M. Radzikowska, J. Simpson, and K. Uszkalo. 2011. Data Mining with Criminal Intent: Final White Paper.

[11] P. Dent. 2000. "Ego-Surfing" Derides Valid, Prudent Activity. In *Online Journalism Review, USC Annenberg School for Communication*.

[12] C. Dummitt. 2017. *Unbuttoned: A History of Mackenzie King's Secret Life.* McGill-Queen's University Press.

[13] C. Ess. 2006. Ethical Pluralism and Global Information Ethics. *Ethics and Information Technology* 8, 4 (2006), 215–226.

[14] C. Fiesler and N. Proferes. 2018. "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society* 4, 1 (2018).

[15] C. Frankel, M. Swain, and V. Athitsos. 1996. *WebSeer: An Image Search Engine for the World Wide Web.* Technical Report 96-14. University of Chicago.

[16] S. Graham, I. Milligan, and S. Weingart. 2015. *Exploring Big Historical Data: The Historian's Macroscope.* Imperial College Press.

[17] B. Hallinan, J. Brubaker, and C. Fiesler. 2019. Unexpected Expectations: Public Reaction to the Facebook Emotional Contagion Study. *New Media & Society* (2019).

[18] S. High. 2015. *Oral History at the Crossroads: Sharing Life Stories of Survival and Displacement.* UBC Press.

[19] H. Jenkins. 2006. *Convergence Culture: Where Old and New Media Collide.* NYU Press.

[20] B. Jules, E. Summers, and V. Mitchell. 2018. Ethical Considerations for Archiving Social Media Content Generated by Contemporary Social Movements: Challenges, Opportunities, and Recommendations.

[21] K. Keats-Rohan (Ed.). 2007. *Prosopography Approaches and Applications: A Handbook.* Oxford.

[22] M. Kirschenbaum, R. Ovenden, and G. Redwine. 2010. *Digital Forensics and Born-Digital Content in Cultural Heritage Collections.* CLIR Publication No. 149. Council on Library and Information Resources.

[23] E. Klein, B. Alex, C. Grover, C. Coates, A. Quigley, U. Hinrichs, J. Reid, N. Osborne, and I. Fieldhouse. 2014. Trading Consequences: Final White Paper.

[24] Y. LeCun, Y. Bengio, and G. Hinton. 2015. Deep Learning. *Nature* 521, 7553 (2015), 436–444.

[25] J. Lin, I. Milligan, J. Wiebe, and A. Zhou. 2017. Warcbase: Scalable Analytics Infrastructure for Exploring Web Archives. *ACM Journal on Computing and Cultural Heritage* 10, 4 (2017), Article 22.

[26] S. Lomborg. 2013. Personal Internet Archives and Ethics. *Research Ethics* 9, 1 (2013), 20–31.

[27] A. Markham and E. Buchanan. 2012. Ethical Decision-Making and Internet Research: Recommendations from the AOIR Ethics Working Committee (Version 2.0).

[28] D. Maron, D. Berry, F. Payton, S. Lakin, and E. White. 2017. "Can We Really Show This"? Ethics, Representation and Social Justice in Sensitive Digital Space. In *JCDL*. 354–355.

[29] S. McTavish. 2018. West Hollywood Goes Global: Exploring Queer Identity on GeoCities. In *Global Digital Humanities Symposium*.

[30] J. Metcalf and K. Crawford. 2016. Where Are Human Subjects in Big Data Research? The Emerging Ethics Divide. *Big Data & Society* 3, 1 (2016).

[31] I. Milligan. 2016. Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives. *International Journal of Humanities and Arts Computing* 10, 1 (2016), 78–94.

[32] I. Milligan. 2017. Welcome to the Web: The Online Community of GeoCities and the Early Years of the World Wide Web. In *The Web as History*, N. Brügger and R. Schroeder (Eds.). UCL Press.

[33] I. Milligan. 2019. GeoCities. In *SAGE Handbook of Web History*, Niels Brügger and Ian Milligan (Eds.). SAGE Publications.

[34] I. Milligan. 2019. *History in the Age of Abundance? How the Web is Transforming Historical Research.* McGill-Queen's University Press.

[35] F. Moretti. 2007. *Graphs, Maps, Trees: Abstract Models for Literary History.* Verso.

[36] J. Nicholas. 2014. A Debt to the Dead? Ethics, Photography, History, and the Study of Freakery. *Histoire sociale/Social history* 47, 93 (2014), 139–155.

[37] H. Nissenbaum. 2011. A Contextual Approach to Privacy Online. *Daedalus* 140, 4 (2011), 32–48.

[38] K. Ocamb. 2012. David Bohnett: Social Change through Community Commitment. *Frontiers* (2012).

[39] Office of the Secretary of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1979. *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research.* Technical Report. Department of Health, Education, and Welfare.

[40] Interagency Advisory Panel on Research Ethics. 2018. Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans — TCPS 2.

[41] R. Peterson. 1997. Eight Internet Search Engines Compared. *First Monday* 2, 2 (1997).

[42] L. Putnam. 2016. The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast. *The American Historical Review* 121, 2 (2016), 377–402.

[43] R. Rosenzweig. 2003. Scarcity or Abundance? Preserving the Past in a Digital Era. *The American Historical Review* 108, 3 (2003), 735–762.

[44] N. Ruest, J. Lin, I. Milligan, and S. Fritz. 2020. The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives. *arXiv:2001.05399* (2020).

[45] B. Sawyer and D. Greely. 1999. *Creating GeoCities Web Sites.* Muska and Lipman Publishing.

[46] J. Smith, J. Burgoyne, I. Fujinaga, D. De Roure, and J. Downie. 2011. Design and Creation of a Large-Scale Database of Structural Annotations. In *ISMIR*. 555–560.

[47] M. Smith. 1999. Invisible Crowds in Cyberspace. In *Communities in Cyberspace*, M. Smith and P. Kollock (Eds.). Psychology Press.

[48] J. Vitak, K. Shilton, and Z. Ashktorab. 2016. Beyond the Belmont Principles: Ethical Challenges, Practices, and Beliefs in the Online Data Research Community. In *CSCW*. 939–951.

[49] H. Webb, M. Jirotka, B. Stahl, W. Housley, A. Edwards, M. Williams, R. Procter, O. Rana, and P. Burnap. 2017. The Ethical Challenges of Publishing Twitter Data for Research Dissemination. In *WebSci*. 339–348.

[50] H. Yang, L. Liu, I. Milligan, N. Ruest, and J. Lin. 2019. Scalable Content-Based Analysis of Images in Web Archives with TensorFlow and the Archives Unleashed Toolkit. In *JCDL*. 436–437.

[51] M. Zimmer. 2018. Addressing Conceptual Gaps in Big Data Research Ethics: An Application of Contextual Integrity. *Social Media + Society* 4, 2 (2018).

[52] M. Zimmer. 2018. How Contextual Integrity can help us with Research Ethics in Pervasive Data. https://medium.com/pervade-team/how-contextual-integrity-can-help-us-with-research-ethics-in-pervasive-data-ef633c974cc1